(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 July 2001 (26.07.2001)

PCT

(10) International Publication Number
WO 01/54110 A1

(51) International Patent Classification[7]: G09G 5/00, 5/08, G06K 9/36, G03B 21/26

(21) International Application Number: PCT/US01/01583

(22) International Filing Date: 18 January 2001 (18.01.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/176,534 18 January 2000 (18.01.2000) US

(71) Applicant *(for all designated States except US)*: **THE TRUSTEES OF THE UNIVERSITY OF PENNSYLVANIA** [US/US]; 3700 Market Street, Suite 200, Philadephia, PA 19104-3147 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: **TAYLOR, Camillo, J.** [JM/US]; 429 West Upsal Street, Philadelphia, PA 19119 (US).

(74) Agent: **McCONATHY, Evelyn, H.**; Dilworth Paxson LLP, 3200 Mellon Bank Center, 1735 Market Street, Philadelphia, PA 19103-7595 (US).

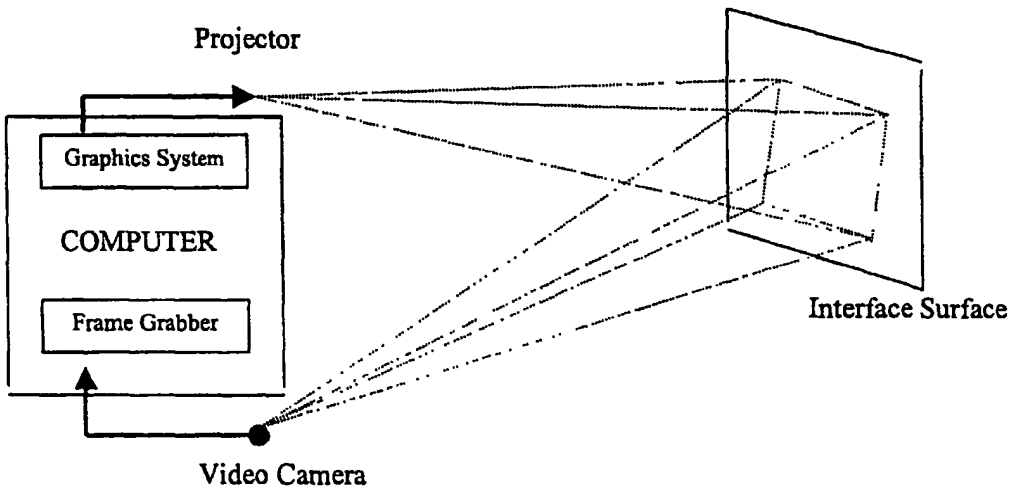
(81) Designated States *(national)*: AU, CA, JP, US.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: VISION-BASED HUMAN COMPUTER INTERFACE SYSTEM



(57) Abstract: This invention provides a novel approach to vision-based human computer interaction in which traditional input and output devices, e.g., monitors, keyboards, touch screens, and mice, are replaced with augmented reality displays (the Augmented Reality display in Fig. 5), projection systems (the projector in Fig.1) and cameras (the camera in Fig. 1). User input is accomplished by projecting an image of the interface onto a flat surface (the Interface Surface in Fig. 1), which is monitored with a video camera (the camera in Fig. 1). The relationship between the three surfaces of interest, i.e., the work surface, the virtual keyboard and the image obtained by the camera, can be characterized by projective transformations of RP^2, which leads to a fast and accurate online calibration algorithm.

WO 01/54110 A1

Vision-Based Human Computer Interface System

REFERENCE TO RELATED APPLICATIONS

This application claims priority to U.S. Provisional Application 60/176,534 filed January 18, 2000.

GOVERNMENT SUPPORT

This work was supported in part by a grant from the National Science Foundation (NSF) Proposal No. 0083240.

BACKGROUND OF THE INVENTION

This invention relates generally to the field of vision-based human computer interface interactions

Vision based interface ideas were proposed by Krueger (*Artificial Reality 2*, Addison-Wesley, Reading, Mass, 1991; *Communications of the ACM*, 36(7):36-38 (1993); U.S. Pat. No. 4,843,568), who described a number of vision-based human computer interface systems, including VIDEOPLACE and VIDEODESK. In these systems, one or more fixed cameras were used to observe the action of the operator and to interpret his/her intentions. The user was observed against a known background, and a shadow obtained by simple background subtraction was composited with an image of the interface and presented to the user in a display. The user was able thereby able to control the interface simply by moving his/her shadow around on the screen.

A similar system, called the Digital Desk, was described by Wellner, *Communications of the ACM* 36(7):86-96 (1993)), in which a projection system was used to project the image of an interface onto a work surface which was monitored by an overhead camera. The video imagery obtained through this camera was analyzed by the computer and correlated with the signal obtained from a microphone mounted on the table to determine when the user touched various interface elements on the display.

However, neither of these systems offers sufficient functionality to provide a practical application because neither observation has been captured into a fast and accurate online, real-time calibration algorithm. Thus, they do not permit acceleration or

augmentation of the reality displays, nor can the systems in the prior art compensate for changes in the relationship between the camera and the interface surface, which occur when either the camera or the interface surface is moved.

Ramesh *et al.*, In *Rendering Techniques 98, Proceedings of the 9th EuroGraphics Rendering Workshop* (1998a), and In *SIGGRAPH* (1998b), describe their visions of the office of the future, which would be equipped with multiple camera and projector systems for achieving spatially immersive user interfaces. They describe techniques for implementing multi projector, multi-surface immersive displays and for acquiring information about the 3D structure of the environment using structured light and stereo techniques. Their work extends the research done on immersive virtual reality systems, exemplified by the CAVE system (Cruz-Niera *et al.*, In *SIGGRAPH*, 1993), and the interface scheme could effect limited interactions between the user and the computer system.

Saund and his colleagues at Xerox PARC (Black *et al.*, In *AAAI Spring Symposium on Intelligent Environments*, 1-6 (1998)) developed a whiteboard scanning system, called the ZombieBoard. This system is able to interpret certain markings on a whiteboard, such as check marks and buttons, and to decipher a limited range of gestures made by the user. Cipolla *et al.*, *J. Image and Vision Computing*, 14(3):171-178 (1996) also refer to vision-based human computer interfaces. Recent work by AT&T labs has tracked colored markers for human computer interfaces, and Microsoft Research lab U.K. has tracked the movement of human hands (MacCormick *et al.*, European Conference on Computer Vision, pgs. 3-(19 June 2000). However, none of these works adequately explore the use of vision-based display systems to provide real-time feedback to the user during interaction.

Thus, there has remained a need in the art for the vision based interaction techniques provided by the present invention, in which no mechanical input devices, such as keyboards, mice and touch screens, are needed, and there is no physical instantiation of the interface. Such a system would provide a previously unavailable level of abstraction, which can be exploited in terms of significantly enhanced flexibility to specify the layout and action of the user interface entirely in software without being constrained by a fixed mechanical interface, to customize interfaces to individual user-defined requirements and capabilities, and to permit the scale-up or miniaturization of the interfaces in ways that cannot be matched in fixed size, monitor based systems.

There is, of course, a large and growing body of literature devoted to the problem of tracking human motion in video imagery, including recent work on the automatic interpretation of sign language by a computer (Starner *et al.*, In *IEEE Trans. Pattern Anal. Machine Intell.* 20(12) (1998); Vogler *et al.*, In *International Conference on Computer Vision* (1998); Ju *et al*, In *Proc. IEEE Conf. on Comp. Vision and Pattern Recog.*, 595-601 (1997); Black *et al.*, In *Motion-Based Recognition*, 245-269 (Shah and Jain, eds.), Kluwer Academic Publishers, Boston, 1997). However, the problems associated with interpreting the users motion in 3D space extend beyond the solution presented by the present invention, which is restricted to monitoring the user's interaction with a 2D surface.

Moreover, Smith *et al.*, *IEEE Computer Graphics and Applications* 18(3):54-60 (1998) described a human computer interface scheme in which a projection system is used to project an image of the interface onto a work surface. However, in that system the user's interaction with the surface is detected by monitoring the electromagnetic field in and around the work surface with field sensors. However, such a system differs from that of the present invention, which provides purely vision-based interface schemes.

Summary of the Invention

The present invention embodies a novel system, method for its use, and article by which a user interacts with a vision-based human computer, in which traditional input and output devices, *e.g.*, monitors, keyboards and mice, are replaced with augmented reality displays, projection systems and cameras. User input is accomplished by projecting an image of the interface onto a flat surface, which is monitored with a video camera. The relationship between the three surfaces of interest, the work surface, the virtual keyboard and the image obtained by the camera, is characterized by projective transformations of $\mathbb{R}P^2$. This observation leads to a fast and accurate online calibration algorithm.

In an embodied system and method for its use, imaging of the interface display interactively comprises a standard personal computer system; a projector attached to a VGA output port; and an image capturing system. In an alternative system and method of use, the image capture system and interface surface interact in an augmented reality display, wherein the projective transformations are computed from projective transformations of real projective plane $\mathbb{R}P^2$ based upon a set of fiducial markings on the interface surface.

In an embodied article, imaging of the interface display interactively comprises a computer-readable signal-bearing medium; means in the medium for specifying a virtual user interface without physical instantiation; and means in the medium for characterizing relationship interaction between work surface, virtual keyboard and projected image of the virtual keyboard by projective transformations of real projective plane $\mathbb{R}P^2$. In an embodied article, imaging of the interface display further comprises means in the medium for projecting an image, wherein said projector is attached to a VGA output port; and means for capturing said projected image. In an alternative article, the means for image capture system and interface surface interact in an augmented reality display, wherein the projective transformations are computed from projective transformations of real projective plane $\mathbb{R}P^2$ which is based upon a set of fiducial markings on the interface surface.

The projective transformations are computed from at least four distinct, non-colinear point correspondences between frame and image buffers. Thus, the interface characteristics, such as size, color, position and layout, are highly flexible, and subject to reconfiguration by the user. Substantially any smooth, flat surface onto which a projected image can be visualized may be used as the interface surface.

An advantage of the vision based interaction technique of the invention is that it requires no mechanical input devices, such as keyboards, mice or touch screens. There are no moving parts and no wires to connect to the interface surface. By avoiding a physical instantiation of the interface, a level of abstraction is gained which can be exploited in a number of ways. The system designer is given the flexibility to specify the layout and action of the user interface entirely in software, without being constrained by a fixed mechanical interface. Thus, interfaces can be customized to the requirements and capabilities of individual users.

In addition, the article and system are very amendable to miniaturization, thereby permitting interesting applications in the field of wearable computer systems. Moreover, the same article and system can be scaled up or down to very large or very small interfaces, a degree of flexibility that cannot be matched by monitor based systems, which are restricted to a fixed size.

Additional objects, advantages and novel features of the invention will be set forth in part in the description, examples and figures which follow, and in part will become

apparent to those skilled in the art on examination of the following, or may be learned by practice of the invention.

DESCRIPTION OF THE DRAWINGS

The foregoing summary, as well as the following detailed description of the invention, will be better understood when read in conjunction with the appended drawings. For the purpose of illustrating the invention, there are shown in the drawings, certain embodiment(s), which are presently preferred. It should be understood, however, that the invention is not limited to the precise arrangements and instrumentalities shown.

FIG. 1 is a schematic diagram of the components of the projector based interface scheme.

FIG. 2 is a block diagram of the projector based interface system.

FIGs. 3A and 3B depict images of the virtual keyboard. FIG. 3A depicts the frame buffer containing the image of the virtual calculator keypad that is projected onto the interface surface. FIG. 3B depicts the image of the interface surface acquired with the video camera.

FIG. 4 is a diagram showing how projective transformations relate corresponding points on the virtual keyboard with those on the interface surface and the image buffer.

FIG. 5 is a block diagram of the augmented reality interface system.

FIGs. 6A-F depicts both the images acquired by the video camera (FIGs. 6A, 6B and 6C), and the corresponding augmented reality displays produced by the system (FIGs. 6D, 6E and 6F). The user is able to select one of the three shapes for display by "pressing" the corresponding button. A square shape is depicted in FIGs. 6A and 6D; a cross shape is depicted in FIGs. 6B and 6E, and a triangle shape is depicted in FIGs. 6C and 6F.

DESCRIPTION OF PREFERRED EMBODIMENTS OF THE INVENTION

The present invention provides systems and articles, and method for using same, by which techniques related to computer vision and augmented reality are employed to develop novel vision-based human computer interfaces with significantly greater flexibility and functionality, in which traditional input and output devices, monitors, keyboards and mice, are replaced with augmented reality displays, projection systems and cameras. There are no moving parts and no wires to connect to the interface surface. By avoiding a physical

instantiation of the interface, a level of abstraction is gained which can be exploited in a number of ways.

Unlike the vision based interface schemes mentioned in the prior art, the invention exploits the fact that the relationship between the three surfaces of interest, the work surface, the virtual keyboard and the image obtained by the camera, can be characterized by projective transformations of $\mathbb{R}P^2$. This observation leads to a fast and accurate online calibration algorithm. The availability of such a real-time, online calibration scheme opens the way for the use of augmented reality displays, in which the image of the interface is composited with the video imagery. In this situation the calibration system is used to compensate for changes in the relationships between the camera and the interface surface of the types which occur when either the camera or the interface surface is moved. In addition, commonly available graphics accelerators are used to expedite some of the image manipulation operations required by the present interface scheme, so that real-time performance can be achieved on standard PCs.

The systems or articles provide flexibility to the designer allowing the layout and action of the user interface to be specified entirely in software, without being constrained by a fixed mechanical interface. This flexibility permits the interfaces to be customized to the requirements and capabilities of the individual user. Just as a graphical user interface can be programmed to present a number of different interfaces on the same computer, the present invention permits the user to arbitrarily reconfigure the interface.

The size, color, position and layout of the interface elements can all be changed in software to reflect individual needs and tastes. Different interfaces are employed for different tasks, in the same way that different GUI's are presented for different programs.

This characteristic of the interface scheme makes it particularly useful to individuals with special needs, since interfaces are easily tailored to suit the capabilities of each individual user. For example, the invention permits interfaces to be individually developed for users who suffer from repetitive stress disorders, such as carpel tunnel syndrome, which be caused or exacerbated from the inflexible arrangement of standard interface devices.

Moreover, the scheme is very amenable to either scale-up or scale-down to very large or very small interfaces, providing a degree of flexibility that cannot be matched by monitor based systems, which are restricted to a fixed size. The interface scheme is

particularly amenable to miniaturization, which makes possible a variety of interesting applications in the field of wearable computer systems.

In a preferred embodiment of the invention, an image of a calculator keypad can be projected onto a screen as diagramed in FIG. 1. Then by analyzing the images of the screen acquired with a video camera, the system is able to determine what the user is indicating on the virtual keyboard, as depicted in FIG. 2, and to respond appropriately. Such a system is outlined schematically in FIG. 3. Effectively, the projector and camera systems acting in concert form a feedback system in which user interaction is effected by occluding various parts of the projected image.

In another preferred embodiment, the system presents to the user an augmented reality display in which an image of the virtual keyboard is overlaid onto a region of the interface surface. A block diagram of the augmented reality interface system is presented in FIG. 5. Once again the users intent is inferred by monitoring the image obtained by the video camera.

One of the more intriguing applications of the presently embodied interface technology is for creating wearable computer systems, wherein premiums are placed on size and weight, such as those described by Mann, *Technology Review* 102(3):36 (1999); Starner *et al.*, *IEEE Trans. Pattern Anal. Machine Intell.*, 20(12) (1998); Picard, *Proceedings of the IEEE*, 86(8) (1998), and Picard *et al.*, *Personal Technologies* 1:231-240 (1997). For instance, the augmented reality display described in Example 2 is particularly suited to implementation on a head-mounted display, which could be contained in a pair of glasses. A miniaturized camera could be mounted on the glasses in such a way that the video imagery closely approximates the user's viewpoint. Inventories could be quickly assessed while walking through a store. Ultimately fully functional computers with sophisticated interface capabilities, will be small enough to easily carry in a pocket.

Clearly, such a system or article will have many applications, for home, business, commercial or industrial settings; however, it could be invaluable to a physically handicapped individual. The projection based Virtual Keyboard system has already been adapted and tested by the inventor's laboratory for use on a 'smart wheelchair." The interface allowed the user to make selections from virtual buttons projected onto the tray on the wheelchair. As a result, since the system can be customized for a particular individuals abilities in software, it has already proven

advantageous over current systems based on keyboards or other physical selection devices.

The interface surface can be any substantially smooth, flat surface, preferably white or nearly white in color. If the surface can be used to clearly view a projected photograph or overhead transparency, it can also be successfully used for the images related to the present invention. However, the greater the range or number of imperfections or color variations in the interface surface, the greater the variations will be in the information transmitted to the computer from the user. Although variations can be managed by corresponding, calibration algorithmic solutions, which may be developed by one of ordinary skill in the art without undue experimentation, less variation in the interface surface will permit reliable projective transformations. In other words, the greater the reliability of the projected image on the interface surface, the greater the reliability of the position of each point on the screen to the coordinates of its projection on the video image.

The interface surface could be a simple, smooth sheet of white paper, such as the one shown in FIG. 6, comprising a pattern of fiducial marks that are readily recognized and tracked in the video imagery. In this way, the user is presented with an interface to the computing system, like the one shown in FIG. 6, whenever he/she looks at the interface surface. Thus, the virtual keyboard appears registered to the interface surface, creating the illusion of a keypad without the need for a physical interface.

One advantage of such a scheme is that the entire input/output system is contained on the headset, thereby eliminating the need for bulky keyboards, display systems, cables or wiring. Consequently, the computer itself can be miniaturized to the size of a cellular telephone.

Another advantage of using modulated light as an input mechanism is that the system is essentially independent of scale. For example, in the projector based system, described in greater detail in Example 1, the interface can be made as large or as small as needed, by simply changing the relative positions of (distance between) the projector and the interface surface. The same system can be used to place the interface on a sheet of paper, on the surface of a drafting table, or on an entire wall. This capacity is particularly useful and advantageous in immersive virtual reality environments, because a designer would be able to place interface elements on any convenient, suitable surface in the environment. Virtual light switches can be projected onto walls, virtual telephone keypads can be projected onto table tops, and virtual displays can be projected onto desktops.

One of the intriguing advantages presented by the embodied interface scheme is the ability to leverage the considerable bandwidth available in the video signal to implement more sophisticated interfaces than are currently possible with a traditional keyboard and mouse system. The typical computer keyboard contains approximately 100 keys, but only one keycode can be transmitted to the computer at a time. By comparison, a single video image contains roughly a quarter of a million pixels, and all of the intensity measurements are acquired in parallel.

Although not intended to be limiting, one approach to exploiting this bandwidth is by designing interfaces in which the user is presented with a large variety of symbols from which to select by occluding different combinations of regions on the interface, or even on each 'button' itself. For example, it would be quite simple to implement a virtual keyboard containing 10 key regions (one for each finger), wherein the user would be able to select from 1024 different symbols by covering and uncovering various, different combinations of keys. This could be visualized in terms of the interface presented by the keyboard of an organ, from which a musician is able to invoke a wide range of sounds by depressing different sets of keys, the components of which are not mutually exclusive of each other. Such a human computer interaction system might be particularly useful to persons with a limited range of motion, since subtle variations in the pattern of occlusion on a virtual keyboard caused by small motions, can be indexed into a vocabulary of thousands of symbols.

Thus, vision based human computer interaction systems and articles are presented in which the user indicates his/her intention by occluding or disoccluding portions of an interface surface, as exemplified by the following prototype systems, one of which uses a standard computer projection system and another which presents an augmented reality display to the user.

EXAMPLES

Embodiments of the invention is further described in the following examples. These examples are provided for purposes of illustration only, and are not intended to be limiting unless otherwise specified. The various scenarios are relevant for many practical situations, and are intended to be merely exemplary to those skilled in the art. These examples are not to be construed as limiting the scope of the appended claims. Thus, the invention should in

no way be construed as being limited to the following example, but rather, should be construed to encompass any and all variations which become evident in light of the teaching provided herein.

Example 1 – Projection-Based System to Display an Image of the Interface.

As shown in FIG. 1, the setup for the prototype implementation of the first preferred embodiment of the vision-based interaction system comprises 3 primary elements to display an image of the interface to the user: (i) a standard personal computer system, (ii) a projector, which is attached to the VGA output port, and (iii) an image capture system. FIG. 2 is a block diagram of the computational system that underlies the projector based user interface scheme. FIG. 3 depicts the image of a virtual calculator keypad. In FIG. 3A the interface stored in the frame buffer is projected onto the screen, while in FIG. 3B the image of the screen is acquired with a video camera. Coordinate frames of reference can be attached to the frame buffer, the image buffer and/or the interface surface in the usual manner.

The interface scheme hinges on the observation that the relationship between the coordinates of corresponding points in these three frames of reference can be expressed quite elegantly using basic projective geometry. By definition the coordinates of a point in the frame buffer, (x_f, y_f), and the coordinates of its image on the screen of the virtual keyboard, (x_s, y_s), are related by a projective transformation. See, FIG. 4. This relationship is expressed algebraically as:

$$(x_s\, y_s\, 1)^T \propto H_{sf}(x_f\, y_f\, 1)^T \qquad \text{(Equation 1)}$$

wherein $H_{sf} \in GL$ (see, Black *et al.*, In *Motion-Based Recognition*, 245-269 (1997), *supra*).

Similarly, another projective transformation relates the positions of points on the screen to the coordinates of their projections on the video image, (x_i, y_i). Therefore, the relationship between points in the frame buffer and their correspondents in the image buffer are expressed algebraically as:

$$(x_i\, y_i\, 1)^T \propto H_{if}\,(x_f\, y_f\, 1)^T \qquad \text{(Equation 2)}$$

wherein $H_{if} \propto H_{is}\, H_{sf}$.

It is well known that a projective transformation is completely specified if its operation is known on a set of points which constitute a projective basis for the relevant projective space (in this case, the real projective plane $\mathbb{R}P^2$). This suggests a

straightforward calibration scheme for determining the mapping between the frame and image buffers. By simply choosing four distinguished points in the frame buffer, such that no three are co-linear, and then locating their correspondents in the image buffer, the straightforward computation of the projective transformation, H_{if}, is computed from the four point correspondences using standard techniques, *e.g.*, Faugeras, *Three-Dimensional Computer Vision*, MIT Press, 1993.

For completeness, a description of how projective transformations are computed from point correspondences is provided as follows:

Let $(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4)$ denote the homogeneous coordinates of four projective points that form a projective basis for $\mathbb{R}P^2$. A projective transformation is constructed, represented by a matrix $H \in GL$ ((see, Black *et al.*, In *Motion-Based Recognition*, 245-269 (1997), *supra*) that maps the standard projective basis

$(\mathbf{e}_1 = 1\ 0\ 0)^T, \mathbf{e}_2 = (0\ 1\ 0)^T, \mathbf{e}_3 = (0\ 0\ 1)^T, \mathbf{e}_4 = (1\ 1\ 1)^T)$ onto $\mathbf{p}_1 - \mathbf{p}_4$ as follows:

$$H \propto (\lambda_1 \mathbf{p}_1 \ \lambda_2 \mathbf{p}_2 \ \lambda_3 \mathbf{p}_3) \qquad \text{(Equation 3)}$$

where:

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \propto (\mathbf{p}_1\ \mathbf{p}_2\ \mathbf{p}_3)^{-1}\mathbf{p}_4 \qquad \text{(Equation 4).}$$

By construction, this matrix H has the property that $\mathbf{p}_i \propto H\, e_i$ for all i. Given two sets of points, $\mathbf{p}_i$ and $\mathbf{q}_i$, a projective transformation, H, can be constructed, which maps $\mathbf{p}_i$ onto $\mathbf{q}_i$ by constructing the projective transformations, H_1 and H_2, that map the standard basis onto $\mathbf{p}_i$ and $\mathbf{q}_i$, respectively, and then composing the transformations as follows:

$$H \propto H_2\, H_1^{-1} \qquad \text{(Equation 5).}$$

Note that this projective transformation can be computed without any a'priori knowledge of the intrinsic parameters of the camera, or of the geometric relationships between the projector, the interface surface and the camera. The system can be made to calibrate itself automatically by projecting fixed patterns on the screen, which can be recognized and localized in the imagery acquired with the video camera.

If fiducial marks on the interface surface are available, one can compute the projective transformation between the image buffer and the interface surface in a similar manner. This projective transformation can be used to apply a 'keystoning correction' to the frame buffer, so that the projection of the virtual keyboard is properly aligned with the interface surface. In the OpenGL graphics pipeline, this keystoning correction can be implemented quite efficiently by manipulating the projection and viewing matrices appropriately.

Once the estimate for the projective transformation between the frame buffer containing the image of the interface and the image buffer has been obtained, it is used to apply a projective rectification to the video imagery obtained by the camera. This projective rectification is accomplished, for example, quite expeditiously by utilizing the known texture mapping capabilities of modern graphics accelerators. These systems can be programmed to perform arbitrary projective transformations on the image buffer at frame rate without burdening the system CPU.

Once the image has been rectified in this way, or by any other recognized method, the system detects the user's interaction with the surface by analyzing the differences between the image of the virtual keyboard in its frame buffer and the rectified image. During the calibration phase, the system constructs a mapping between color or intensity values projected onto the interface surface and the corresponding color or intensity value that is measured by the camera system. From this mapping the system is able to determine when a particular pixel in the rectified image differs significantly from the expected color or intensity value. The end result of this analysis is a binary image where pixels that differ significantly from their expected values are marked with a 1.

In the special case of a static interface, in which the button elements do not change, this binary image is constructed by simply computing the difference between the current image and a fixed background image.

Once this binary image has been constructed, the system interprets the user's intent by analyzing the pattern of occlusion in the image. For example, the Virtual Calculator system (the embodiment wherein the virtual keyboard is a calculator keypad as shown in FIG. 3), each of the virtual keys is divided into two regions as shown. A "keypress" is detected when the central region is sufficiently occluded, while the peripheral region is left untouched. This scheme allows the system to distinguish the situation in which the user is

simply reaching over one key to point to another, since in this case both regions of the virtual keys that the user is reaching over will be fully or partially occluded.

Many other schemes for analyzing the pattern of occlusion in the image to discern the user's intent are also possible to those skilled in the art. For freehand drawing programs one could simply track the "uppermost" occluded point in the image, and take that as the point of interaction. One could also distinguish between 'intended' and 'unintended' keypresses by instantiating a second 'button,' which the user could occlude once the hand was positioned in the desired location to 'press' the intended button.

In the preferred embodiment, there is an implicit assumption that most of the intensity changes in the video image are due to user interaction. This would not be the case in environments in which there are significant changes in the ambient lighting, or in which shadows are cast over time. Such variables can be ameliorated by employing more sophisticated change detection algorithms which are better able to distinguish between those changes in the image that are due to the users intervention, and those that arise from other unintentional variations. For example, the system could, as an alternative embodiment, be made to adapt to changes in the ambient lighting conditions. In another alternative, it could also employ other cues, such as shape and motion, to improve the detection of the user's hand position.

Another more challenging problem can occur when the system is presented with ambiguous occlusion patterns. For example, it would be difficult to implement standard qwerty or Dvorak style keyboards for touch typing using the preferred embodiment, since it would be difficult to distinguish, solely on the basis of the imagery acquired with a monocular camera system, the situations in which the user's fingers are hovering above the keys, from the situation in which a keypress is actually intended. However, such problems can be avoided by redesigning the interface so that keypresses are indicated by slight motions of the fingers, which cover and uncover parts of the keypad.

Example 2 - Augmented Reality System for Presenting an Image of the Interface

Another preferred embodiment of the vision-based interaction system uses an augmented reality display to present an image of the interface to the user. A block diagram of the augmented reality system is provided in Figure 5. This system is similar to the

system described in Example 1, except the projector has been replaced by an augmented reality display.

Significantly, as opposed to the projection system, in this system the relationship between the camera and the interface surface is allowed to change over time. This means that for every image in the video sequence the system must recompute the projective transformation between the interface surface and the image buffer. One way in which this can be accomplished is by tracking the position of a set of fiducial markings on the interface surface in the video imagery, and performing the computation described above in Equations 3, 4 and 5, wherein projective transformations are computed from point correspondences.

Once the projective transformation has been calculated, it is used to produce an augmented reality display where an image of the virtual keyboard is composited with the video image, so that the interface appears in the correct position on the video image. This technique is described in more detail by Kutulakos and Vallino, *IEEE Transactions on Visualization and Computer Graphics* 4(1):1-21 (1998).

The differences between corresponding images acquired by the video camera and by the augmented reality displays produced by the system are shown in FIGs. 6A through F. FIGs. 6A-6C present the images obtained with the video camera. FIGs. 6D-6F present the augmented reality displays provided to the user. As shown in FIG. 6, the user is given the opportunity to select one of the three shapes for display (a square in FIGs. 6A and 6D; a cross in FIGs. 6B and 6E; or a triangle in FIGs. 6C and 6F) by "pressing" the corresponding button.

The projective transformation is also used to apply a projective rectification to the region of the video imagery that corresponds to the interface surface. Then, the rectified image is analyzed to determine the user's interaction.

Note that in this case, the task of computing a binary image, which indicates that portions of the interface are occluded, is straightforward since it simply amounts to locating dark objects against a lighter background. Such a task is well within the capability of one skilled in the art using standard modern computer vision techniques.

Once this binary image has been computed, the user's intention can be inferred by analyzing the pattern of occlusion. In the system shown in FIG. 6, the system effectively detected which buttons were 'pressed,' *i.e.*, which shapes were selected, based on which shape was occluded, and the appropriate pattern was then displayed in the selection box,

proving the present vision-based human computer interaction system to be useful, reliable and effective.

Each and every patent, patent application and publication that is cited in the foregoing specification is herein incorporated by reference in its entirety.

While the foregoing specification has been described with regard to certain preferred embodiments, and many details have been set forth for the purpose of illustration, it will be apparent to those skilled in the art that the invention may be subject to various modifications and additional embodiments, and that certain of the details described herein can be varied considerably without departing from the spirit and scope of the invention. Such modifications, equivalent variations and additional embodiments are also intended to fall within the scope of the appended claims.

I claim:

1. A system by which a user interacts with a vision-based human computer comprising:
   means for specifying a virtual user interface without physical instantiation; and
   means for characterizing relationship interaction between work surface, virtual keyboard and projected image of the virtual keyboard by projective transformations of real projective plane $\mathbb{R}P^2$.

2. The system of claim 1, wherein the projective transformations are computed from at least four distinct, non-colinear point correspondences between frame and image buffers.

3. The system of claim 1, wherein interface characteristics are subject to reconfiguration by the user, wherein said characteristics may be selected from the group consisting of size, color, position and layout.

4. The system of claim 3, wherein the computer is sufficiently miniaturized as to permit the computer to be worn by the user as clothing or apparel.

5. The system of claim 1, wherein the interface surface is any substantially smooth, flat surface onto which a projected image can be visualized.

6. The system of claim 1, wherein interactive imaging of the interface display comprises:
   a standard personal computer system;
   a projector attached to a VGA output port; and
   an image capturing system.

7. The system of claim 5, wherein the image capture system and interface surface interact in an augmented reality display, and wherein the projective transformations are computed from projective transformations of real projective plane $\mathbb{R}P^2$, which is based upon a set of fiducial markings on the interface surface.

8. A method by which a user interacts with a vision-based human computer comprising specifying a virtual user interface without physical instantiation.

9. The method of claim 8, wherein layout and action of the user interface is specified entirely by software.

10. The method of claim 8, wherein relationship interaction between work surface, virtual keyboard and projected image of the virtual keyboard is characterized by projective transformations of real projective plane $\mathbb{R}P^2$.

11. The method of claim 10, wherein the projective transformations are computed from at least four distinct, non-colinear point correspondences between frame and image buffers.

12. The method of claim 8, wherein interface characteristics are subject to reconfiguration by the user, and wherein said characteristics may be selected from the group consisting of size, color, position and layout.

13. The method of claim 11, wherein the computer is sufficiently miniaturized as to permit the computer to be worn by the user as clothing or apparel.

14. The method of claim 8, wherein the interface surface is any substantially smooth, flat surface onto which a projected image can be visualized.

15. The method of claim 8, wherein imaging of the interface display interactively comprises:
    a standard personal computer system;
    a projector attached to a VGA output port; and
    an image capturing system.

16. The method of claim 15, wherein the image capture system and interface surface interact in an augmented reality display, and wherein the projective transformations are computed from projective transformations of real projective plane $\mathbb{R}P^2$ based upon a set of fiducial markings on the interface surface.

17. An article by which a user interacts with a vision-based human computer comprising:
    a computer-readable signal-bearing medium;
    means in the medium for specifying a virtual user interface without physical instantiation; and
    means in the medium for characterizing relationship interaction between work surface, virtual keyboard and projected image of the virtual keyboard by projective transformations of real projective plane $\mathbb{R}P^2$.

18. The article of claim 17, wherein the projective transformations are computed from at least four distinct, non-colinear point correspondences between frame and image buffers.

19. The article of claim 17, wherein interface characteristics are subject to reconfiguration by the user, wherein said characteristics may be selected from the group consisting of size, color, position and layout.

20. The article of claim 19, wherein the computer is sufficiently miniaturized as to permit the computer to be worn by the user as clothing or apparel.

21. The article of claim 17, wherein the interface surface is any substantially smooth, flat surface onto which a projected image can be visualized.

22. The article of claim 17, wherein imaging of the interface display further comprises:

means in the medium for projecting an image, wherein said projector is attached to a VGA output port; and

means for capturing said projected image.

23. The article of claim 22, wherein the means for capturing said image and interface surface interact in an augmented reality display, and wherein the projective transformations are computed from projective transformations of real projective plane $\mathbb{R}P^2$, which is based upon a set of fiducial markings on the interface surface.

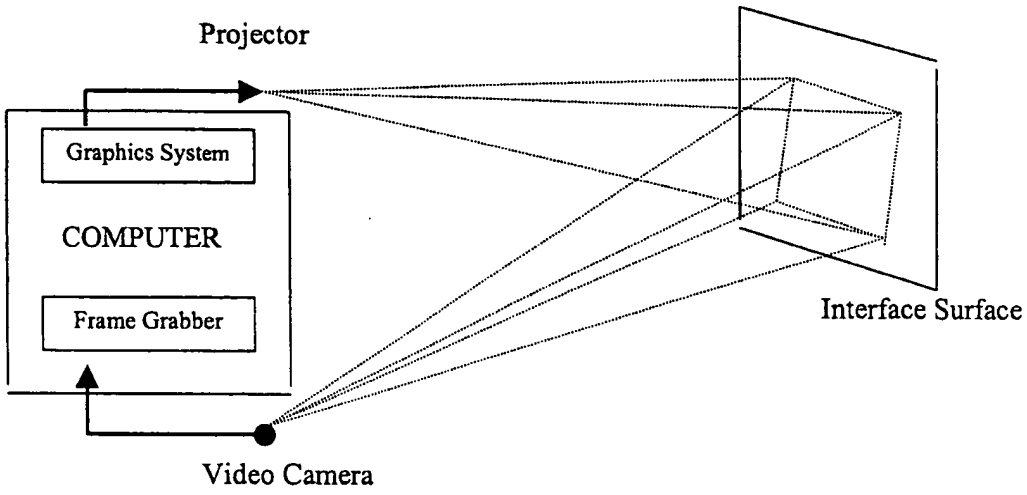

Projector
Graphics System
COMPUTER
Frame Grabber
Interface Surface
Video Camera

FIG. 1

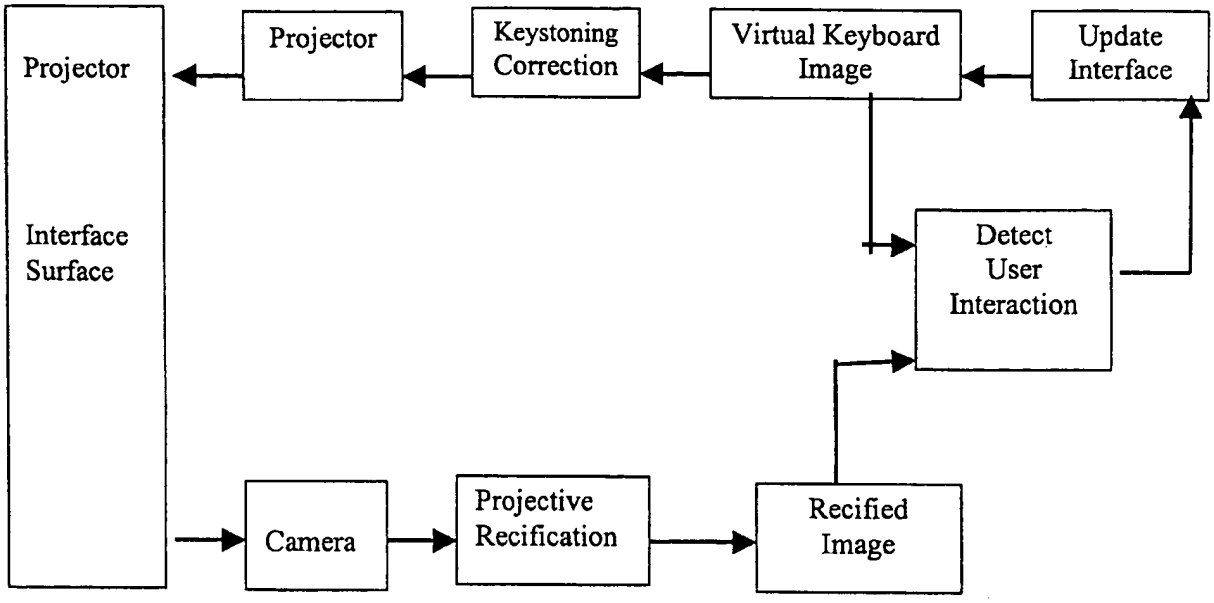
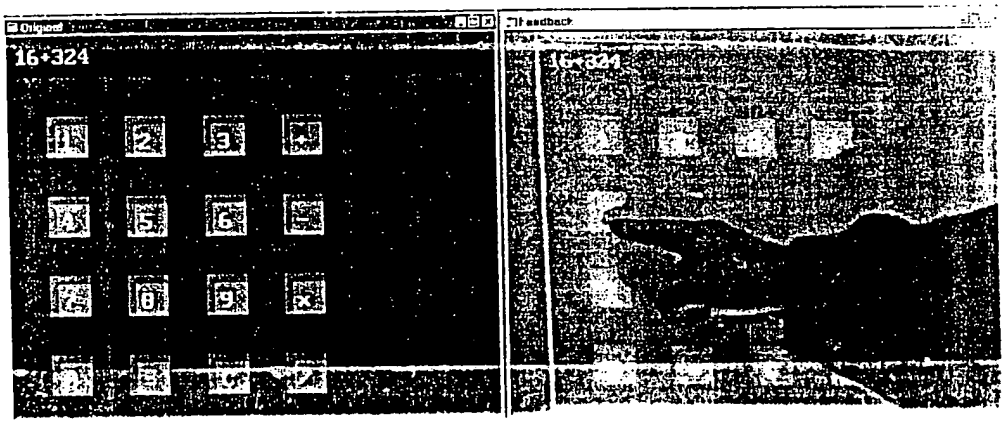

Projector
Interface Surface
Projector
Keystoning Correction
Virtual Keyboard Image
Update Interface
Detect User Interaction
Camera
Projective Recification
Recified Image

FIG. 2



A B

FIG. 3

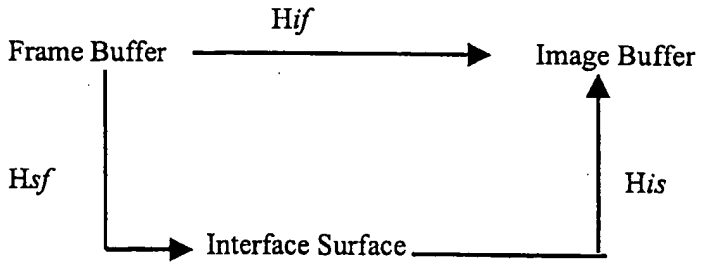


FIG. 4

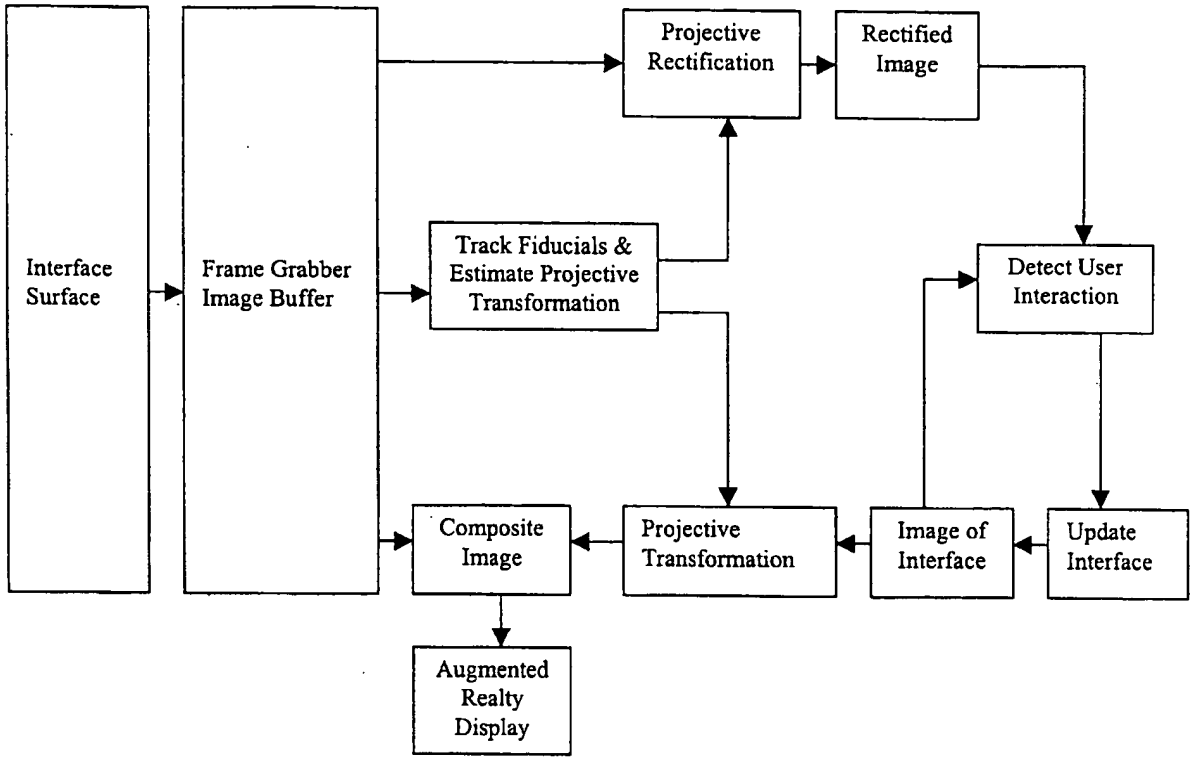

Interface Surface
Frame Grabber Image Buffer
Projective Rectification
Rectified Image
Track Fiducials & Estimate Projective Transformation
Detect User Interaction
Composite Image
Projective Transformation
Image of Interface
Update Interface
Augmented Realty Display

FIG. 5

FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/01583

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : G09G 5/00, 5/08; G06K 9/36; G03B 21/26
US CL : 345/156, 157, 158; 382/276, 291; 353/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 345/156, 157, 158; 382/276, 291; 353/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,528,263 A (PLATZKER et al) 18 June 1996, see column 4 line 44 to column 5 line 62.	8, 9, 12, 14, 15
A	US 6,005,547 A (NEWMAN et al) 21 December 1999, see all	
A, P	US 6,147,678 A (KUMAR et al) 14 November 2000, see all.	
A, E	US 6,198,485 B1 (MACK et al) 06 March 2001, see all.	

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 April 2001 (17.04.2001)	09 MAY 2001
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703)305-3230	Authorized officer Kent Chang Telephone No. 703-305-9700

Form PCT/ISA/210 (second sheet) (July 1998)

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US01/01583

Continuation of Item 4 of the first sheet: The title is not short and precise.

—— NEW TITLE ——

Vision-Based Human Computer Interface System